Automatic Topic Model Adaptation for Sentiment Analysis in Structured Domains

Geoffrey Levine and **Gerald DeJong**
University of Illinois at Urbana-Champaign
Urbana, IL
levine@illinois.edu and mrebl@illinois.edu

Abstract

We present a novel topic modeling approach to sentiment analysis for documents organized into hierarchical categories. In our approach, positive, negative, and subject matter topics are learned and used to infer document labels. A Markov chain Monte Carlo model procedure adapts the number and structure of topics based on a minimum description length objective function. We apply our approach to Yelp.com business reviews and Amazon.com book reviews and demonstrate that 1) the model adaptation procedure selects a high quality model from the space of alternatives, and 2) the resulting model performs well relative to state of the art regression and topic modeling approaches.

1 Introduction

Selecting an appropriate model is an important part of any machine learning endeavor. The model must be chosen in a manner so as to balance two objectives: 1) Be sufficiently rich to capture the relevant patterns in the data, and 2) Be simple enough to avoid spurious patterns in the training data (overfitting). In natural language processing tasks, there are often many modeling choices to be made regarding what feature granularities and interactions to consider. It is important to make these decisions in a manner such that the resulting model strikes a balance between these two somewhat contradictory objectives.

In order to appropriately make these choices, we must consider not only the task involved but also the training data available. With copious data we can reliably calibrate complex models, but with limited data complex models risk overfitting. Many general model selection techniques exist in which each candidate model is fit to the training data and scored with respect to a particular criterion. While these approaches allow us to compare a small number of models in order to select the most appropriate, they require calibrating each model's parameters to the training data. However, when there are many modeling choices to be made and thus a large space of alternative models, fitting all of them to the training data is computationally prohibitive.

In this paper, we present a novel topic modeling approach for structured sentiment analysis domains and an automatic model adaptation approach that takes advantage of categorical metadata. This model adaptation approach resolves the structure of the metadata with the significant patterns in the training data to determine the number and range of latent topics.

We demonstrate our approach on Yelp.com business reviews as well as Amazon.com book reviews. We show that our model adaptation approach selects an appropriate model given a particular amount of training data, and the resulting model is high quality relative to alternative regression and topic modeling approaches.

2 Background

Sentiment analysis, in which the opinion of the author is estimated from a document, has recently grown in popularity. Many works have explored unigram models (Pang and Lee, 2005; Snyder and Barzilay, 2007). Higher-order n-gram models are explored in (Pang and Lee, 2008; Baccianella et al., 2009). In order to combat the high dimensional feature space that accompanies such models, models restricting features based on part of speech patterns (Baccianella et al., 2009) or opinion templates (root, modifiers, negation words) (Qu et al., 2010) have been introduced.

Topic models are generative models in which the words in a document are assumed to be asso-

Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 75–83,
Chiang Mai, Thailand, November 8 – 13, 2011.

ciated with one of a number of abstract "topics." Latent Dirichlet allocation (Blei et al., 2003) is a popular topic model in which the topic distribution per document is assumed to have a Dirichlet prior. In supervised LDA (Blei and McAuliffe, 2007), the distribution of document topics is used to produce a document label. (Zhao et al., 2010; Titov and McDonald, 2008b; Titov and McDonald, 2008a) focus on topic modeling based approaches to aspect-based sentiment summarization, identifying product features and the opinion associated with each.

Model selection is the act of using data to choose a statistical model from a set of candidates. Often, this task is performed by fitting each candidate model to the training data and using a criterion to score the models and select one. Popular criteria include the Akaike information criterion (AIC) (Akaike, 1974), the Bayesian information criterion (BIC) (Schwarz, 1978), and the minimum description length principle (MDL) (Rissanen, 1978; Grunwald, 2007). Structural Risk Minimization (Vapnik, 1995) defines a general framework in which a nested hierarchy of hypotheses can be defined based on prior knowledge of the domain, such that a hypothesis balancing goodness of fit with simplicity can be identified. The work presented in this paper is closely related to the model adaptation procedure presented in (Levine et al., 2010), in which a hill-climbing approach is used to explore a large space of generative models.

3 Topic Modeling for Sentiment Analysis in Structured Domains

Our approach takes advantage of hierarchical categorical metadata. Formally, this hierarchy forms a tree structure, which we refer to as $\mathcal{C}$ (See Figure 1). Individual nodes in the tree are called *categories*, for which we use notation c. A *categorization*, $\mathbf{c}$, is a set of categories, $\mathbf{c} = \{c_{d,1}, c_{d,2}, ...\}$. $\mathbf{c}$ can be thought of as metadata about a product/service being reviewed. For example, with regard to a book review, $\mathbf{c}$ could equal {"Fiction", "Fiction\Drama", "Fiction\Drama\Romance", ... }. $\mathbf{c}$ must be well formed. That is, if a node $c \in \mathcal{C}$ appears in categorization $\mathbf{c}$, all ancestors of c (in the tree $\mathcal{C}$) must also appear in $\mathbf{c}$. $\mathbf{c}$ can contain multiple distantly related categories. For example, a particular book could belong to both "Fiction\Poetry" and "Children\Humor."

Documents, or examples, are denoted $d = \langle \mathbf{x_d}, \mathbf{c_d}, y_d \rangle$. $\mathbf{x_d} = [w_{d,1}, w_{d,2}, ..., w_{d,|\mathbf{x_d}|}]$ is a vector of *words*. Each word is an element from the *vocabulary*, $V = \{w_1, w_2, ..., w_{|V|}\}$. $\mathbf{c_d}$ is the document's categorization. y_d is a numeric *rating* from a discrete space ({1,2,3,4,5} for our domains). The rating is an overall score given by the document's author to the product or service being reviewed.

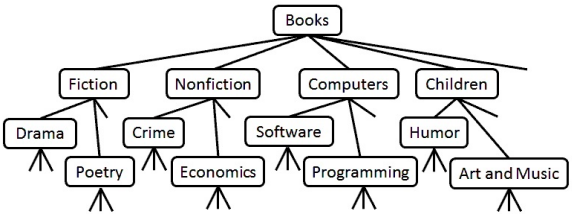

Figure 1: A subtree of the category tree, $\mathcal{C}$, corresponding to the Amazon Books domain.

We are given a collection of documents, $\mathcal{D}$, and our goal is to learn a function $f(\langle \mathbf{x}, \mathbf{c} \rangle)$ to predict rating $\hat{y}$ from an *unlabeled document* so as to minimize the expected loss over the unknown distribution of documents:

$$E\left(\text{loss}(y, f(\langle \mathbf{x}, \mathbf{c} \rangle))\right) \quad (1)$$

We use the squared error loss function.

3.1 Model Structure

We will start by presenting our generative document topic model. In this model, each review is composed of a mixture of topics, and each topic is associated with a distribution over words. We use $t \in T$ to denote a topic, and P_t to denote t's word distribution. Within a document, each word is assumed to be generated from a particular topic, although which topic is unobservable. In many topic model approaches, such as latent Dirichlet allocation (Blei et al., 2003), topics are learned in an unsupervised or weekly supervised fashion (as is the case with supervised LDA).

In our model, we assume each document is generated according to a rigid topic distribution (more similar to labeled LDA (Ramage et al., 2011)). Each document is a mixture of three topics: 1) a *positive* topic (+), in which the reviewer is speaking favorably about the product/service, 2) a *negative* topic (-), in which the reviewer is speaking unfavorably, and 3) a *subject* topic (s_i) corresponding to general text about the content/features of the product.

The proportion of positive words to negative words is a function of the rating score. Subject topics reflect the language used when discussing a particular product or group of products, and do not directly influence a document's rating. Still, learning these topics appropriately is crucial to the performance of the model. When a word is indicative of either the positive or negative topic, it is important to account for its probability in the subject topic. For example, the word "good" may be less indicative of a book review's rating if the review discusses a book about ethics. Furthermore, if subject topics are not learned appropriately, words related to the subject matter of products/services with a disproportionate number of positive training reviews would be attributed to the positive word topic. This will lead to poor performance on unseen data. On the other hand, if these words are correctly attributed to the subject topic, then the high ratings will appropriately be attributed to the unconditional positive words appearing in the reviews.

What constitutes a subject worthy of having its own topic? For books, should we only make broad distinctions such as fiction vs. non-fiction? Should we learn a unique subject topic for each book? Should we use something in between these two extremes? In answering these questions, we need to balance goodness of fit to the training data with model simplicity. There is no optimal answer, it is a function both of the domain (in that we need to make the most "significant" distinctions), and the amount of training data available to calibrate our model (more training data allows us to reliably learn the additional parameters introduced by making additional distinctions).

There exists a many-to-one relationship between documents and subject topics. The mapping from document to subject topic is a function of the document's categorization, $s_i = g(\mathbf{c_d})$, $s_i \in T$. We call the function g the *topic mapping function*. The range of g is the set of subject topics, $\{s_1, s_2, ..., s_N\} \subset T$. In Sections 3.1.1 and 3.1.2 we assume that g is fixed. In Section 3.2, we consider exploring the space of alternative topic mapping functions.

We assume that in expectation, a fixed but unknown fraction, α of each document is composed of the subject topic. The remainder of the review is composed of the positive and negative topics, and the positive/negative ratio is related to the document's rating. Let y_{min} and y_{max} represent the minimum and maximum scores in the rating scale. For document d with score y_d the expected fractional breakdown into topics is as follows:

$$\begin{aligned}
\texttt{Positive:} \quad f_+(y_d) &= (1-\alpha)\frac{y_d - y_{min}}{y_{max} - y_{min}} \\
\texttt{Negative:} \quad f_-(y_d) &= (1-\alpha)\frac{y_{max} - y_d}{y_{max} - y_{min}} \\
\texttt{Subject:} \quad f_s(y_d) &= \alpha
\end{aligned} \tag{2}$$

In total, a model M is composed of the topic mapping function, the value α, and the word distributions associated with each topic. $M = (g, \alpha, P_+, P_-, P_{s_1}, P_{s_2}, ..., P_{s_N})$.

3.1.1 Training

Expectation maximization (Hastie et al., 2001) can be used to train our topic model. The procedure works by iteratively updating 1) the assignment of words in each document to latent topics (Expectation Step), and 2) the word distributions associated with each topic (Maximization Step). EM proceeds as follows:

Expectation Step

Each word is assigned an expected topic distribution. For word i in document d:

$$\begin{aligned}
q_{d,i}(+) &= \frac{f_+(y_d)P_+(w_{d,i})}{Z_{d,i}} \\
q_{d,i}(-) &= \frac{f_-(y_d)P_-(w_{d,i})}{Z_{d,i}} \\
q_{d,i}(g(\mathbf{c_d})) &= \frac{f_s(y_d)P_{g(\mathbf{c_d})}(w_{d,i})}{Z_{d,i}} \\
Z_{d,i} &= \sum_{t \in \{+,-,g(\mathbf{c_d})\}} f_t(y_d)P_t(w_{d,i})
\end{aligned} \tag{3}$$

Maximization Step

Topic word distributions are updated so as to maximize the likelihood of the training data. For each topic t:

$$P_t(w) = \frac{\sum_{d \in \mathcal{D}} \sum_{i=1}^{|\mathbf{x_d}|} \mathbf{I}_w(w_{d,i}) q_{d,i}(t)}{\sum_{d \in \mathcal{D}} \sum_{i=1}^{|\mathbf{x_d}|} q_{d,i}(t)} \tag{4}$$

where

$$\mathbf{I}_w(w_{d,i}) = \begin{cases} 1 & \text{if } w_{d,i} = w \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

3.1.2 Inference

Given the trained topic models we use Bayes' theorem to compute the probability that an unlabeled document $\langle \mathbf{x_d}, \mathbf{c_d} \rangle$ is associated with a particular rating. Let $T_d = \{+, -, g(\mathbf{c_d})\}$:

$$P(y|\mathbf{x_d}, \mathbf{c_d}) = \frac{P(y)P(\mathbf{x_d}, \mathbf{c_d}|y)}{P(\mathbf{x_d}, \mathbf{c_d})}$$
$$= \frac{P(y)\prod_{i=1}^{|\mathbf{x_d}|}\left(\sum_{t\in T_d} f_t(y)P_t(w_{d,i})\right)}{\sum_{y'} P(y')\prod_{i=1}^{|\mathbf{x_d}|}\left(\sum_{t\in T_d} f_t(y)P_t(w_{d,i})\right)} \quad (6)$$

For evaluation purposes, we output the expected value of y and compute the squared error to the true value.

3.2 Model Adaptation

In this section we introduce a Markov chain Monte Carlo approach to selecting a topic mapping function g. Here, we stochastically explore the space of topic mapping functions, driven by the minimum description length principle and estimates of the effect of modifications to g. This approach resists local minima and efficiently finds a high quality topic mapping function.

3.2.1 Minimum Description Length Objective

Our goal is to find a model that balances fit to the training data with simplicity, and concentrates its flexibility where most useful to capture relevant patterns in the domain. We accomplish this by utilizing a two part minimum description length objective function. The objective is the sum of 1) the description length (in bits) required to encode the model and 2) the description length of the data given the model.

$$\mathrm{L}(M, \mathcal{D}) = \mathrm{L}(M) + \mathrm{L}(\mathcal{D}|M) \quad (7)$$

where $\mathrm{L}(M)$ is a function of the number of model parameters ($\approx$ the product of the number of topics and the vocabulary size) and $\mathrm{L}(\mathcal{D}|M)$ is the negative log likelihood of the data given the model. Thus the goal is to jointly minimize the complexity of the model and maximize the likelihood of the data given the model, and the objective can be rewritten as:

$$\mathrm{L}(M, \mathcal{D}) = \beta(N+2)|V| + -log\left(l(\mathcal{D}|M)\right) \quad (8)$$

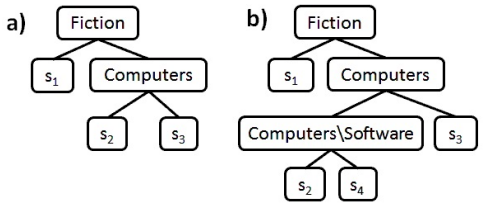

Figure 2: Two possible partitioning trees for the Amazon.com Books category tree (Figure 1). Tree b) is formed by splitting s_2 in a) based on membership in the "Computer\Software" category.

where β is a complexity penalty constant, which is selected via cross-validation.

3.2.2 Topic Mapping Functions

The topic mapping function g maps from categorization $\mathbf{c}$ to a subject topic $s_i \in T$. We select a particular g from the space of *binary partitioning trees*, G. In a binary partitioning tree, each internal node references a category c, and each leaf node references a subject topic s_i. See Figure 2. Starting at the root, a categorization, $\mathbf{c}$ is recursively assigned by each internal node to 1) the left subtree if the referenced category c is in $\mathbf{c}$, and 2) the right subtree otherwise, until a leaf (with associated subject topic) is reached. For example, within the book review domain, a node may reference the category "Computers." In this case, computer books are recursively assigned a subject topic by the left subtree, and all others by the right subtree. We allow only well formed partitioning trees: Any node in g that references a category c with parent category $parent(c) \in \mathcal{C}$ must have an ancestor that references $parent(c)$. For example, we do not allow a node in g to reference "Computers\Software", unless we have already conditioned on the "Computers" category. This constraint guarantees that we partition the space of categorizations into coherent regions (we would never assign "Computer\Hardware" and "Fiction" books to the same subject topic while assigning "Computer\Software" to a different topic).

3.2.3 Adjacent Model Estimation

In order to guide the search through G, we consider 2 types of modification operations: We can 1) Split a leaf based on category $c \in \mathcal{C}$, splitting one partition into two, adding an additional subject topic to the model, or 2) Merge two leaves with the same parent, combining two partitions into one, removing a subject topic from the model.

Given a particular g, there are a finite number of possible merge and split operations to the leaves. Key to our search is the fact that we can estimate the change to the description length objective that each possible modification will cause, using the latent topic distributions assigned to each topic during expectation maximization. Consider merging two subject topics s_i and s_j:

$$\widehat{\Delta L} = -\beta|V| - \sum_{t\in\{s_i,s_j\}}\sum_{w\in V} \#_t(w) log\frac{P_{m_{i,j}}(w)}{P_t(w)} \tag{9}$$

where

$$\#_t(w) = \sum_{d\in\mathcal{D}}\sum_{i=1}^{|\mathbf{x_d}|}\mathbf{I}_w(w_{d,i})q_{d,i}(t)$$

$$P_{m_{i,j}}(w) = \frac{\sum_{t\in\{s_i,s_j\}}\#_t(w)}{\sum_{t\in\{s_i,s_j\}}\sum_{d\in(D)}\sum_{i=1}^{|\mathbf{x_d}|}q_{d,i}(t)} \tag{10}$$

Now consider splitting subject topic s_i based on category c:

$$\widehat{\Delta L} = \beta|V| - \sum_{t\in\{s_{i,c},s_{i,!c}\}}\sum_{w\in V} \#_t(w) log\frac{P_t(w)}{P_{s_i}(w)} \tag{11}$$

where

$$\#_{s_{i,c}}(w) = \sum_{\substack{d=(\mathbf{x},\mathbf{c},y)\in\mathcal{D}\\ s.t. c\in\mathbf{c}}}\sum_{i=1}^{|\mathbf{x_d}|}\mathbf{I}_w(w_{d,i})q_{d,i}(s_i)$$

$$P_{s_{i,c}}(w) = \frac{\#_{s_{i,c}}(w)}{\sum_{\substack{d=(\mathbf{x},\mathbf{c},y)\in\mathcal{D}\\ s.t. c\in\mathbf{c}}}\sum_{i=1}^{|\mathbf{x_d}|}q_{d,i}(s_i)}$$

$$\#_{s_{i,!c}}(w) = \sum_{\substack{d=(\mathbf{x},\mathbf{c},y)\in\mathcal{D}\\ s.t. c\notin\mathbf{c}}}\sum_{i=1}^{|\mathbf{x_d}|}\mathbf{I}_w(w_{d,i})q_{d,i}(s_i)$$

$$P_{s_{i,!c}}(w) = \frac{\#_{s_{i,!c}}(w)}{\sum_{\substack{d=(\mathbf{x},\mathbf{c},y)\in\mathcal{D}\\ s.t. c\notin\mathbf{c}}}\sum_{i=1}^{|\mathbf{x_d}|}q_{d,i}(s_i)} \tag{12}$$

These estimates are upper bounds on the change to the description length objective function. Incorporating these changes (and the associated word distributions) and then retraining the model with expectation maximization may further reduce the objective. These bounds serve as a guide to estimate the objective for models that have *not* been fit to the training data, which will drive our search through G for the optimal topic mapping function.

3.2.4 MCMC exploration

Markov chain Monte Carlo (Gilks et al., 1996) stochastically steps through the space of alternative topic mapping functions. At each iteration of MCMC, the topic model with the current topic mapping function is fit to the training data and the objective change associated with all possible merges and splits is estimated. We then construct a proposal distribution for alternative models that can be reached with these operations. Limiting the proposal distribution to these candidate models, as in (Titov and Klementiev, 2011) and (Singh et al., 2011) induces a decomposable, feasible computation. A model is sampled from this distribution and adopted if certain criteria on its fitness are met.

MCMC will converge to a probability distribution over models. By making better models (those with a lower objective) more probable, the MCMC chain will be driven towards higher quality models. We use an exponential distribution over models:

$$P(M) = \frac{e^{-L(M,\mathcal{D})}}{Z_P} \tag{13}$$

with normalization factor Z_P.

The proposal distribution, Q assigns some probability to all candidate models that can be reached by a single merge or split to each of the leaves in the current partitioning tree. In Q, splits and merges to leaves without a common parent are independent by construction. Now, consider a particular leaf, l, that has the following possible splits, $S = \{c_1, c_2, ..., c_l\}$, and cannot be merged with another leaf. For example, in Figure 2a, the leaf corresponding to s_1 meets this condition as it cannot be merged to another leaf and has possible splits {"Fiction\Drama", "Fiction\Poetry", "Nonfiction", "Computers", "Children", ... }. Let M_l represent the subset of models where l is not split, and M_{l,c_i} represent the subset of candidate models where l has been split with respect to category c_i.

$$Q(M_l) = \frac{e^{-\tau L(M,\mathcal{D})}}{Z_l}$$

$$Q(M_{l,c_i}) = \frac{e^{-\tau \widehat{L}(M_{l,c_i},\mathcal{D})}}{|S|Z_l}$$

$$Z_l = \left(e^{-\tau L(M,\mathcal{D})} + \sum_{c\in S}\frac{1}{|S|}e^{-\widehat{L}(M_{l,c},\mathcal{D})}\right) \tag{14}$$

$0 < \tau \leq 1$ controls a balance between having the proposal distribution completely influenced by

the estimated objectives vs. a uniform proposal distribution.

For all pairs of leaves that share a common parent, we entertain a merge operation. In Figure 2a the leaves corresponding to s_2 and s_3 meet this criteria. Suppose two leaves, l and l' have possible splits $S = \{c_1, c_2, ..., c_l\}$ and $S' = \{c'_1, c'_2, ..., c'_{l'}\}$ respectively. In addition to the one merged alternative, there are $w = (|S|+1)(|S'|+1)$ alternatives that involve only splits to the two leaves. Let $M_{l,l'}$ represent the subset of candidate models where l and l' are merged

$$Q(M_{l,l'}) = \frac{w^{-\frac{1}{2}} e^{-\tau \widehat{L}(M_{l,l'},\mathcal{D})}}{w^{-\frac{1}{2}} e^{-\tau \widehat{L}(M_{l,l'},\mathcal{D})} + (Z_l)(Z_{l'})} \quad (15)$$

The factor of $w^{-\frac{1}{2}}$ accounts for the difference between the number of neighbors that the models with l and l' merged vs. split have. If the two leaves are not merged, then the conditional probability for each of the $(|S|+1)(|S'|+1)$ remaining structural alternatives is computed in Equation 14.

A new topic mapping function g' is sampled from Q and fit to the training data via the expectation maximization presented in Section 3.1.1. If a random value sampled uniformly from $\mathcal{U}[0, 1)$ is less than

$$\frac{P(M_{g'})Q(g|g')}{P(M_g)Q(g'|g)} \quad (16)$$

then g' is accepted as the new topic mapping function g^{t+1}. This guarantees that the Markov chain will converge to the distribution P as $t \to \infty$. Because the ratio $P(M_{g'})/P(M_g)$ appears in Equation 16, the normalization factor Z_P in Equation 13 cancels out and does not need to be computed.

4 Empirical Evaluation

We perform a set of experiments to demonstrate the following:

1) Given the topic model structure outlined in Section 3.1, the model adaptation procedure in section 3.2 selects a high performing topic mapping function while only evaluating a small fraction of the total number of funtions.

2) The topic model resulting from model adaptation is high quality relative to alternative state-of-the-art approaches.

4.1 Data

We demonstrate our approach to two structured sentiment analysis datasets. First, we gathered a collection of approximately 8,000 Yelp.com business reviews from the greater New York area. For this data, businesses are assigned into categories and subcategories based on the Yelp.com business hierarchy. There are 22 primary categories {Arts and Entertainment, Education, Financial Services, Restaurants,...}, each with 6 to 100 subcategories (restaurants have the most subcategories, {Japanese, Barbeque, Cafe, Fast Food, Burgers, Ultra High Enc, Formal, Full Bar,...}). Businesses can be assigned to multiple categories and subcategories within the hierarchy.

Second, we utilize 20,000 Amazon.com book reviews, extracted from the data set first presented in (Qu et al., 2010). Categorical distinctions in these domains are related to the Amazon.com product hierarchy. A small portion of the product hierarchy appears in Figure 1. Books can be assigned to multiple distantly related categories. For example, the book *Six Wives of Henry VIII* belongs to "History\Europe\England\Tudor and Stuart," "Biographies and Memoirs\Specific Groups\Women" and three other categories

For each domain, we have at most one review corresponding to any particular business/book. This allows for a broad coverage of the space of categorizations.

4.2 Results and Discussion

To compensate for extreme variations in the training data we apply two smoothing steps. First, we found that for longer reviews, the assumption that each word is drawn independently from the document's topics is too strong, and so for reviews with more than 35 words, we scale the term counts such that the total is 35. Second, because of the large size of the vocabulary, after training, some rare words have zero or near zero probability in some of the topics. When these words are observed during inference, they have a very strong effect on the document's expected rating. We found that smoothing the subject topics with the overall word distribution across topics stabilizes the predicted ratings and improves performance. The amount of smoothing could be optimized to maximize the likelihood of the test data, but we found that performance varied little for a wide range of values and so we choose a 1 to 1 smoothing.

From each dataset, we sample a subset of size 1000 for cross validation parameter tuning and use the remaining examples for experimentation.

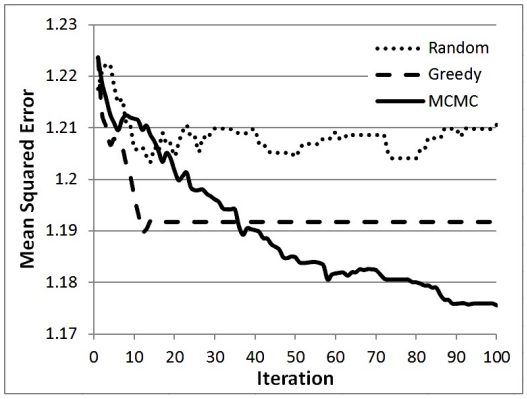

Figure 3: Learning curves for three model sampling approaches on Yelp.com test data with 500 training examples (averaged over 20 trials).

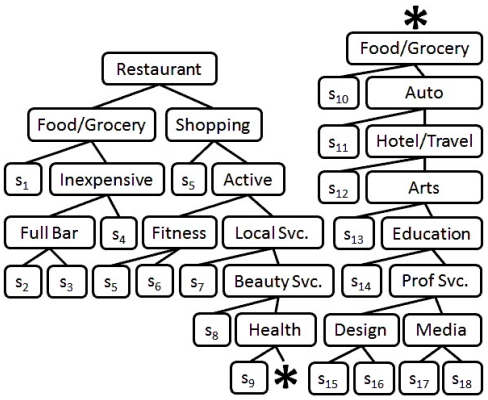

Figure 4: A representative partitioning tree learned from 500 training examples on the Yelp.com data.

The validation data is used to learn the values of α, the subject topic fraction, and β, the complexity penalty. We found that setting τ, the MCMC smoothing factor, equal to .1 worked well across our datasets. For each trial, then, disjoint training and test sets are sampled from the remaining data.

First, we apply our Markov chain Monte Carlo model adaptation procedure along with greedy and random alternatives to demonstrate the necessity of a directed and stochastic approach. In the greedy approach, at each iteration we estimate the objective for all candidate models that can be reached with a single split/merge to each subject topic and adopt the model with the minimum estimate. For the random approach, at each iteration we start with the simplest topic mapping function (mapping all categorizations to one subject topic), and uniformly at random add distinctions until the model has the same number of subject topics as the optimal model found by the MCMC approach. We choose this instead of sampling at random from all possible topic mapping functions as the vast majority of such functions have nearly as many subject topics as training examples. For each approach, at iteration i, we chart the test mean squared error for the best (lowest objective) model observed during training in iterations 1 to i.

Figure 3 charts the per iteration mean squared error on the Yelp test data for the three model adaptation approaches. The greedy approach initially makes the fastest progress, but it is susceptible to local minima, and it levels off before being overtaken by the MCMC approach. As the random approach does not leverage the data in determining what distinctions to make, it fails to make progress at the rate of the other approaches. Its poor performance is indicative of the importance of having an efficient directed model adaptation approach, as high performing models are few and far between, even if we limit our search to models of the appropriate complexity level (number of subject topics). Figure 4 shows a representative partitioning tree learned from the Yelp.com dataset.

Next we compare our approach to alternative regression and topic modeling approaches. In order to implement regression, we 1) Form a vector of unigram (and optionally bigram) occurrences normalized to length 1 (which we found to work better than unnormalized or frequency vectors), and 2) Form a vector corresponding to categorical membership with one element for each node in category tree $\mathcal{C}$. For each example, we set each element in the vector to value γ if the example belongs to the corresponding category, and 0 otherwise. The feature vector is the concatenation of these two vectors. We tested three regression approaches: ridge regression, lasso, and ϵ-support vector regression with a quadratic kernel (Chang and Lin, 2001). In each case, the cross validation dataset is used to tune the value of γ and the regularization parameter (for ridge regression and lasso) or ϵ and the cost parameter (for ϵSVR). We found that in all cases, lasso and ϵSVR were ouperformed by ridge regression, and so omit their results.

For the supervised latent Dirichlet allocation approach, as the space of labels is numeric and discrete, we can treat the task either as a regression problem (Blei and McAuliffe, 2007), or as a multiclass classification problem (Wang et al., 2009).

	Yelp.com Data					Amazon.com Data				
	Training Examples					Training Examples				
	500	1000	2000	4000	6000	500	1000	2000	4000	6000
TMSD, MCMC	**1.207**	**1.062**	**.983**	**.915**	**.870**	**1.243**	**1.161**	**1.090**	**1.019**	**.981**
TMSD, Simple	1.252	1.108	1.017	.951	.893	1.300	1.256	1.158	1.075	1.027
TMSD, Complex	1.284	1.123	—-	—-	—-	1.281	1.198	—-	—-	—-
RR, Uni	1.319	1.182	1.103	1.020	.949	1.337	1.265	1.145	1.081	1.033
RR, Uni/Bi	1.285	1.164	1.059	.971	.903	1.310	1.237	1.119	1.041	1.001
SLDA	1.664	1.649	1.606	1.556	1.479	1.621	1.632	1.607	1.581	1.555

Figure 5: Mean Squared Error for 1) the presented topic model for structured domains (TMSD) using MCMC Model Adaptation (MCMC), the simplest topic mapping function (Simple), or the most complex topic mapping function (Complex), 2) ridge regression (RR) with unigrams (Uni) or unigrams and bigrams (Uni/Bi), and 3) multiclass supervised latent Dirichlet allocation (SLDA). Results are averaged over 10 trials, each with 1000 test examples. The MCMC approach significantly outperforms all other approaches for each training set size (Yelp.com: $p < .01$, Amazon.com: $p < .05$).

We used an open source implementation of each approach, (Chang, 2010) and (Wang, 2009), and found that utilizing the multiclass approach and predicting the expected rating based on the posterior likelihood of each class outperformed the regression approach, so we present these results. The cross validation data is used to learn the number of latent topics and Dirichlet distribution parameter.

For the Markov chain Monte Carlo approach, in order to hasten learning for this comparison, starting from the simplest topic mapping function, we perform a greedy model adaptation until reaching an estimated local minimum, and then apply 50 additional iterations of MCMC model adaptation.

Figure 5 shows the average mean squared error for each approach for various amounts of training data. Our topic model with model adaptation has lower error than each of the alternatives. Paired t-tests reveal that the differences are statistically significant in all cases ($p < .01$ for all Yelp.com and $p < .05$ for all Amazon.com tests). Using MCMC model adaptation also outperforms using either the simplest topic mapping function or the most complex mapping function (which maps each distinct training categorization to a different subject topic).

Ridge regression with unigrams uses the same word and categorical representations as our approach. However, it is unable to entertain the nonlinear relationships between document categorizations and words and is outperformed in all cases. Bigrams improve the performance of ridge regression, especially for larger amounts of training data. This suggests that accounting for word ordering could potentially improve the performance of our topic model as well. sLDA is unable to take advantage of the categorical information during topic construction, and with the limited training data available, its performance is marginally better than guessing the mean label (MSE: 1.675 for Yelp.com data and 1.660 for Amazon.com data).

5 Conclusion

We present an approach to sentiment analysis for structured domains. In our approach, positive, negative, and subject topics are learned and used to infer document labels. Partitioning tree based topic mapping functions define the number and structure of subject topics. A Markov chain Monte Carlo model adaptation procedure explores the space of topic mapping functions based on a minimum description length objective. We demonstrate the approach on two sentiment analysis domains and show that the model adaptation procedure efficiently finds a high performance model that leverages the categorical structure of the documents to outperform other regression and topic modeling approaches.

Acknowledgment

This material is based upon work supported by the Office of Naval Research under Award No. ONR Grant N00014-09-1-0693. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the Office of Naval Research.

References

H. Akaike. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

S. Baccianella, A. Esuli, and F. Sebastiani. 2009. Muli-facet rating of product reviews. In *The 31st European Conference on Information Retrieval Research*, pages 461–472.

D. Blei and J. McAuliffe. 2007. Supervised topic models. In *Neural Information Processing Systems*.

D. Blei, A. Ng, and M. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Chih-Chung Chang and Chih-Jen Lin, 2001. *LIBSVM: A Library for Support Vector Machines*. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

J. Chang. 2010. Lda: Collapsed gibbs sampling methods for topic models. online.

W. Gilks, S. Richardson, and D. Spiegelhalter. 1996. *Markov Chain Monte Carlo in Practice*. Chapman and Hall.

P. Grunwald. 2007. *The Minimum Description Length Principle*. The MIT Press, Cambridge, Mass.

T. Hastie, R. Tibshirani, and J. Friedman. 2001. *The Elements of Statistical Learning*. Springer, New York, NY.

G. Levine, G. DeJong, L. Wang, R. Samdani, S. Vembu, and D. Roth. 2010. Automatic model adaptation for complex structured domains. In *The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, pages 243–258.

B. Pang and L. Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *The 43rd Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, page 124.

B. Pang and L. Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1–2):1–135.

L. Qu, G. Ifrim, and G. Weikum. 2010. The bag-of-opinions method for review rating prediction from sparse text patterms. In *The International Conference on Computational Linguistics*, pages 913–921.

D. Ramage, D. Hall, R. Nallapati, and C. Manning. 2011. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *The Conference on Empirical Methods in Natural Language Processing*.

J. Rissanen. 1978. Modeling by shortest data description. *Automatica*, 14:445–471.

G. E. Schwarz. 1978. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464.

S. Singh, A. Subramanya, F. Pereira, and A. McCallum. 2011. Large-scale cross-document coreference using distributed inference and hierarchical models. In *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*.

B. Snyder and R. Barzilay. 2007. Multiple aspect ranking using the good grief algorithm. In *The 8th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 300–307.

I. Titov and A. Klementiev. 2011. A bayesian model for unsupervised semantic parsing. In *The 49th Annual Meeting of the Association for Computational Linguistics*.

I. Titov and R. McDonald. 2008a. A joint model of text and aspect ratings for sentiment summarization. In *The 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 308–316.

I. Titov and R. McDonald. 2008b. Modeling online reviews with multi-grain topic models. In *The Seventeenth International Conference on World Wide Web*, pages 111–120.

V. Vapnik. 1995. *The Nature of Statistical Learning Theory*. Springer-Verlag.

C. Wang, D. Blei, and L. Fei-Fei. 2009. Simultaneous image classification and annotation. In *IEEE Conference on Computer Vision and Pattern Recognition*.

C. Wang. 2009. Supervised latent dirichlet allocation for classification. online.

W. Zhao, J. Jiang, H. Yan, and X. Li. 2010. Jointly modeling aspects and opinions with a maxent-lda hybrid. In *The Conference on Empirical Methods in Natural Language Processing*, pages 56–65.